Multiple Personalities on the Web: A Study of Shared Mboxes in FOAF

Jennifer Golbeck, Thameem Khan, Nilay Sanghavi, Nishita Thakker

College of Information Studies
University of Maryland, College Park
College Park, MD 20742 USA

Abstract. The Friend-of-a-Friend Vocabulary (FOAF) is used in many online social networks to represent information about users and their friendships. Previous work has looked at how FOAF can be used to merge accounts across social networks. We often consider that merging as a benefit to the user, connecting their personal information and friend lists which would otherwise stay isolated. However, users may create multiple accounts - even on the same network - to intentionally separate their data. FOAF makes it equally easy to merge these accounts. In this paper, we are interested in the impact Semantic Web reasoning - with no additional data mining technology - can be used to resolve multiple accounts and the implications that has for privacy and safety online. We crawled FOAF profiles from all the social networking website that generate it, and looked at the profiles of individuals with multiple accounts to understand how they were using these accounts and why they were created. We present the results of this analysis and discuss the implications.

1 Introduction

Just as social networks are one of the most popular web-based activities FOAF is one of the most widely used ontologies on the Semantic Web. With its semantics and the application of a Semantic Web reasoner, it is possible to merge profiles and friendship connections within and across social networking websites. While this capability is generally viewed as one of if not the greatest powers and benefits of the Semantic Web, it also has implications for privacy and security. Users may want to create multiple profiles to keep parts of their lives separate, and this type of reasoning would impinge on the privacy they expect. At the same time, these multiple accounts could be used for nefarious purposes, and the reasoning and merging may allow us to detect and prevent this bad behavior.

In this paper, we ask one core question: Using Semantic Web reasoning on existing FOAF files, what can we learn about the frequency of, reasons for, and use of multiple profiles belonging to the same person in online social networks? While there are mechanisms for generating FOAF when it is not available and for merging profiles outside of Semantic Web reasoning, our work is a study of this problem on the Semantic Web as it exists today. Thus we are interested

only in Semantic Web technologies and *existing* Semantic Web data. In addition to focusing our results on a specific problem, this restriction means that the conclusions we draw are immediately applicable because they cover existing data analyzed with existing techniques.

2 Related Work

2.1 FOAF Syntax and Semantics

The FOAF vocabulary[1] is used to describe people, their attributes, and the relationships between people. FOAF is written in OWL and takes advantage of some specific OWL semantics. Specifically, the vocabulary employs the Inverse Functional Property on many of the properties connecting a person to an account. This includes `foaf:aimChatID`, `foaf:homepage`, and, most importantly for our purposes, `foaf:mbox` and `mbox:sha1_sum`.

These latter two properties connect a person to their email address, either using the email address directly or using the SHA1 hash of the address respectively. Inverse Functional Properties serve as unique identifiers; in this case, it means only one person can own a given email address. Thus, two people with the same email address can be inferred to be the same person. All the social networks that generate FOAF use the `mbox:sha1_sum` to identify users. The value for this property is a hash of the user's email address in the form `mailto:example@example.com`. Nearly all the networks provide this for every user. A few networks do not require users to provide email addresses; if no email is available, the property is not included in the FOAF output. Datasets are discussed further in section 3.

Because the `mbox:sha1_sum` is an inverse functional property, we can use it to merge accounts. Using an OWL reasoner, all accounts that share a common value for the `mbox:sha1_sum` are inferred to be the same account.

2.2 FOAF in the Wild

Previous research has looked at how FOAF is being used on the web. In 2005, [1] collected a set of FOAF documents and analyzed the commonly used properties and the social network structure within that set. The state of FOAF has certainly changed over the past four years, and the original study did not look comprehensively at the available FOAF but rather used a somewhat arbitrary collection of FOAF documents found by Google, and those that could be reached by crawling from this set.

A later study in 2008 [2] extensively collected FOAF documents from all the social networking website that generate it, and used that to study how extensively Semantic Web reasoning could link accounts in different networks. Those results showed that using the `mbox:sha1_sum` property enables accounts to be merged connecting every pair of networks and that the percentage of nodes that bridge networks are roughly what would be expected in a social network.

[1] http://xmlns.com/foaf/spec/

Table 1. The social networks used in this study and the number of members used in our analysis. Note that we often could not find all members of a given network, so these numbers do not represent the total membership of the sites.

Network	**Purpose**	**Members Studied**
Advogato	Business	2,778
Buzznet	Photos	208,324
DeadJournal	Blogging	9,801
eCademy	Business	61,242
FilmTrust	Social/Entertainment	1,250
GreatestJournal	Blogging	36,862
InsaneJournal	Blogging	1,410
LiveJournal	Blogging	3,563,267
Minilog.com	Blogging	119
Rossia.org	Blogging	4,180
Tribe	Social/Entertainment	218,694

3 Data Sources and Methodology

3.1 Data Sources

We are interested in finding accounts that would be merged by applying OWL reasoning to FOAF data. Unlike previous research that studied the impact this had on connecting social networks, we are not interested in social connections at all. Rather, we want to know the implications that arise from identifying multiple profiles as belonging to the same person. To do this, we used essentially the same data as was used in [2] with slightly expended crawls. In this section, we will explain the datasets in detail.

Eleven social networking websites generate FOAF files for their users and we used all of these networks in our research. Note that this is not just the total number of networks we used, but *all* the web-based social networks with available FOAF. LiveJournal is the largest of those, accounting for just over 75% of the users. All of these networks were crawled in [2] and we used the same dataset. That data was collected in 2008 and while there are certainly more accounts on these websites now, for our purposes of understanding multiple accounts on social networks, the dataset was completely sufficient.

For each network, we gathered as many profiles as possible. Some networks - FilmTrust, Ecademy, and Advogato - provided a full list of all of their members.In the rest of the networks, a full list of members was not available, and thus we had to crawl the network. To do so, we chose several users as starting nodes and performed a breadth first search through the network to find all reachable members. For each user, we accessed the FOAF file, pulled URIs of their friends' FOAF files, and added those URIs to our queue. Table 1 shows the number of users in each network that we were able to use in this study.

There are almost certainly smaller components of these networks that our crawls did not reach. Also, users with no social connections would never be

discovered on a crawl. However, since we are looking for multiple profiles and not examining social connections, missing profiles will not have a significant impact on our results. Furthermore, any applications using FOAF would need to follow the same procedures we did in this study, and thus our data set is representative of what FOAF applications would use. In the worst case, we will underestimate the number of profiles a person has, but our sample should provide representative insights into the implications of merging profiles with FOAF.

For every member we were able to include in the study, we accessed their FOAF file. For the purpose of this work, we were interested only in the member's friends and unique identifiers (given by the inverse functional properties). Thus, to save space and increase efficiency, we implemented a task-specific OWL reasoner that considers only the FOAF inverse functional properties and foaf:knows property, and ignores the rest of the data.

Traditionally, a reasoner would not keep track of the sources of each axiom in the knowledge base. Since we are specifically interested in how data is repeated in multiple sources, we added a provenance tracking feature to our reasoner. This maintains a record of the document where each axiom is asserted. With this data available, it is straightforward to identify on which and how many social networks a member has accounts, as well as the sources for each friendship.

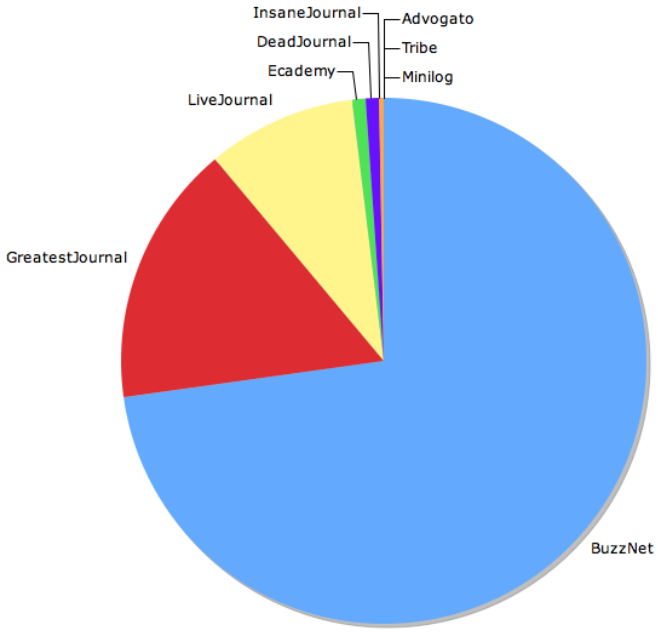
3.2 Frequency of Multiple Accounts

Once we had collected the FOAF files as described above, we implemented a simple customized OWL reasoner that would match up profiles with the same `mbox:sha1_sum`. We ran this over the data and identified individuals with multiple accounts.These multiple accounts occurred both within and across networks, though it was much more common to have multiple accounts within a network.

Of all `mbox:sha1_sum`s used on multiple accounts, 83.6% existed on only one network. Nearly all the rest - 16.0% - were spread between two networks. Among the `mbox:sha1_sum`s on one network, the vast majority are on Buzznet, at 72.9%. The now defunct GreatestJournal had 15.9%, LiveJournal has 9.3%, and the rest of the networks have less than 1% of the users. This latter distribution is shown in figure 3.2.

There were 982,913 unique `mbox:sha1_sum` values among all the datasets. Among these, 47,563 are associated with multiple accounts. The vast majority - 77.6% (36,918) - are used on only two accounts. If we add in the 13.0% that had three associated accounts, this comprises over 90% of the total. Only 16 `mbox:sha1_sum`s had more than 30 associated accounts. The number of accounts associated with each `mbox:sha1_sum` follows a power law distribution (see figure 2). The top ten `mbox:sha1_sum`s with the most associated accounts had 350, 118, 108, 92, 62, 59, 56, 42, 40, and 39. Details about these are discussed in section 4.3 and 4.3.

Fig. 1. Frequency of multiple accounts associated with one `mbox:sha1_sum`s on each network



4 Experiment and Results

4.1 Methodology

In order to gain insights into why multiple accounts were associated with the same `mbox:sha1_sum`, we chose several subsets of accounts to examine. First, we selected the top five email addresses with the greatest number of associated accounts. Because there are so many profiles and attributes in these accounts, we believe these users provide the clearest and most extensive picture of multiple account holding online. Then, because smaller numbers of multiple accounts were more common, we selected 40 `mbox:sha1_sum` values at random that had only two associate accounts and another 40 `mbox:sha1_sum` values with 5 associated accounts.

For each of these users, we looked at the publicly accessible personal information for each of their profiles, including name, age, gender, username, the activity on the account (last active dates, latest blog posts, etc), the content of their blog posts (when available), social networking connections, and other data that was present. We then compared and analyzed this data across the accounts held by each person to develop insights into their online personas.

If two different users entered the same email address, we would expect to find very few similarities in their profiles, if any. However, if we find five accounts associated with one `mbox:sha1_sum` and the profiles of each account have the

same age, gender, hometown, and astrological sign, it is highly unlikely that they belong to five random people. The shared `mbox:sha1_sum` implies that the accounts belong to the same person. In our analysis of accounts, we looked for evidence to support this inference. Only when we found few or no similarities did we conclude that the accounts may belong to different people.

4.2 Hypotheses

We had several theories as to why users would have multiple accounts:

1. Users opened an account and then forgot that it existed or forgot their password. They then opened up another account to replace it. This is easily detectable by looking at the latest activity on the account when it was available.
2. Users create different accounts to compartmentalize parts of their lives. For example, a user may have one account for personal social networking and entertainment use, one for business use, one for religious use, etc. The goal in this case would be to keep these parts of the user's life separate online.
3. Users create accounts for separate topics. Unlike H2 where the accounts are used to present different versions of oneself to different audiences, this hypothesis addresses accounts as an organizational mechanism. For example, a blogger may have many blogs on different topics and create a separate account for each. The user presents the same persona on each blog, but separates topics through multiple accounts.
4. Users maintain completely different personas in different accounts. For example, a user may have a profile of a 15-year-old boy and a 50-year-old woman. Names, locations, interests, and other personal information may also change. The intentions behind these multiple personas vary, but this provides one of the more interesting reasons for creating separate accounts.
5. Sybil Attacks [3] are attacks on systems where users create many accounts to cause some damage. This may be in the form of creating accounts that rate one another highly in order to artificially increase the perceived reliability or quality of each individual. They may also be used for voting, commenting, or otherwise creating a larger presence and chance for being heard. Multiple social networking accounts could be used for this purpose as well.

We do not necessarily expect to see instances of all these theories, and we also believe we may find other unexpected reasons behind multiple accounts.

4.3 Most Frequently Merged

There are insights to be found in looking at `mbox:sha1_sum`s associated with many social networking accounts. Do these accounts actually belong to the same individual? If so, this validates the assumptions behind the FOAF model that treats an email address as a unique identifier. Understanding the purpose of these accounts also can help us understand if FOAF reasoning presents a threat

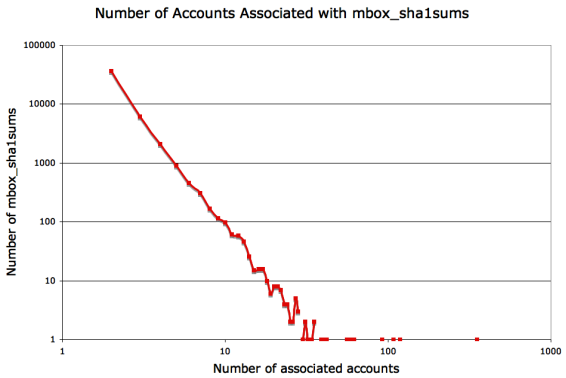


Fig. 2. Number of `mbox:sha1_sums` with a given number of associated accounts. Note that both the x-axis and y-axis are on logarithmic scales.

to privacy that users are trying to maintain, or if it is a feature that can benefit users in maintaining their identities.

We considered the top five most popular `mbox:sha1_sums`.

1. 350 Associated Accounts By far the most common value for the `mbox:sha1_sum` was `08445a31a78661b5c746feff39a9db6e4e2cc5cf`. We found 350 profiles using this address. While it is not possible to easily unhash this address, an educated guess revealed what was happening. The sha1 hash of `mailto:` (i.e. an empty email address) is `08445a31a78661b5c746feff39a9db6e4e2cc5cf`. Thus, when users left their address blank, systems generated this as the value for their `mbox:sha1_sum`.

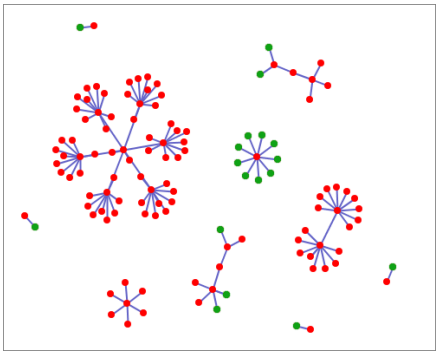
2. 118 Associated Accounts This appears to be a test account of some sort. Ninety of the account names were consecutive integers from 2 through 90 (with 69 and 83 skipped) plus 100 and 1000. The rest of the account names were single letters or common words. The accounts are basically empty of profile information. The only content to be found is collections of photographs in some of the accounts.

The social network of these users supports the hypothesis that they all belong to one user rather than to many different users. Figure 4.3(a) shows the social network for these nodes. Each red node indicates one of the 108 accounts associated with this `mbox:sha1_sum` while green nodes indicate accounts with different `mbox:sha1_sum`. The figure shows all the friendships of these nodes. Note that the vast majority are to other nodes that share this `mbox:sha1_sum`.

3. 108 Associated Accounts This was a very clear case of one user creating multiple accounts. Each profile indicates that it belongs to a 19-year-old female living in a specific Detroit suburb (unnamed here to protect the user's privacy). Each

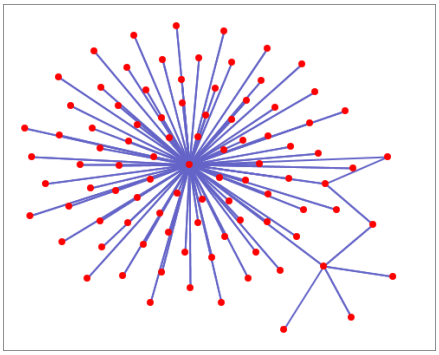
Fig. 3.

(a)



Social network of the 118 accounts sharing one `mbox:sha1_sum`. Green nodes indicate the few accounts that are in this social network but do not share the same email address.

(b)



Social network of the 92 accounts that share one address. Basically all are friends with the central node that shared the same `mbox:sha1_sum`.

has exactly one connection to the same person, a user called"panasonicyouth", a buzznet staff member who is friends with over 575,000 users - the fast majority of people on Buzznet. There is no indication as to what the purpose is for each account - most have no content or posts.

4. 92 Associated Accounts All but three of these accounts were on the Insane Journal blogging website, and the remaining were on Dead Journal. The account names were related to pop music icons, contestants on American Idol, and characters in Disney Channel television series. The profile photos also matched that persona (e.g. a profile with an account name resembling Paris Hilton would also have a picture of Paris Hilton). Forty-nine of the accounts have been shut down by Insane Journal; attempts to access the profiles show the message "This journal has been deleted and purged." The remaining accounts still provide plenty of information about what is happening.

Figure 4.3(b) shows the social network of these accounts, represented as nodes. This shows all the edges for every node. Note that all the accounts are connected, and nearly all nodes have only one connection to the node at the center. That central person shares the same `mbox:sha1_sum` as the other accounts. None of the accounts have social connections to users with a different `mbox:sha1_sum`.

An examination of the remaining accounts reveals their intent. While most accounts have no public posts, one contains a post that explains all of the

acounts. The post states that the author is a high-school student using the accounts to participate in celebrity fan fiction sexual role playing games. The different accounts represent the many different personas the author represents in the games, both male and female. The games take place through private messaging, but the accounts provide a face for the personality in the game. The post explains the author's rules for engaging in a game with someone else and lists characters the student likes to represent (which correspond with the type of celebrities used in the profile names and photos).

Thus, we can conclude that these accounts also belong to the same person who is using each as an alternative identity in role playing games.

5. 62 Associated Accounts The accounts associated with this `mbox:sha1_sum` do not appear to belong to the same person. The profile characteristics vary widely. Profiles are well developed with photos, posts, and discussions. This level of effort makes it unlikely that all the accounts are the work of one individual. It is possible that these users all entered a fake email address that we were not able to identify; the use of fake addresses was common, as we will discuss in section 5.1.

4.4 Five and Two Associated Accounts

We randomly selected 40 `mbox:sha1_sum`s that had five associated profiles. When one or more of the profiles were private or unavailable, we threw out that `mbox:sha1_sum` and randomly selected another. These represented `mbox:sha1_sum`s in the top 5% with respect to the number of associated accounts among `mbox:sha1_sum`s with multiple accounts. We also used the same method to select 40 `mbox:sha1_sum`s that had two associated profiles. Over 75% of the `mbox:sha1_sum`s with multiple accounts had only two, so these are representative of the most common case.

In almost all cases, we found that the accounts associated with a given `mbox:sha1_sum` belonged to the same person. The profile information was largely the same; for example, we often found the same hometown, age, astrological sign, and gender in all the profiles. This, combined with the fact that we know they used the same email address, is extremely strong evidence that all accounts represented the same person. When there were discrepancies, they were often changes in age (e.g. from under 21-years-old to 21) or hometown (e.g. from a small suburban town in one profile to the name of the metropolitan city in another).

Of the 80 total `mbox:sha1_sum`s, we found only four where all the accounts did not obviously belong to the same person - just over 1%.

Interestingly, most of the accounts were essentially unused. They had some basic profile information, but no posts, no friends (beyond, occasionally, the "default" friends on a given network), no photos, and no content of any other type.When at least one account associated with a given `mbox:sha1_sum` has some activity, the others could possibly be mistakes, forgotten, or used for some purpose that cannot be observed. However, it was not infrequent that *all* the

Table 2. A list of fake email addresses and the number of accounts associate with each in our dataset.

123@hotmail.com	14	none@none.com	13	a@hotmail.com	9	a@yahoo.com	9
asdf@hotmail.com	9	me@hotmail.com	8	me@hotmail.com	8	me@yahoo.com	7
123@yahoo.com	6	asdf@yahoo.com	6	none@yahoo.com	6	asd@hotmail.com	5
email@email.com	5	none@hotmail.com	5	123@123.com	4	abc@123.com	4
hot@yahoo.com	4	no@hotmail.com	4	x@hotmail.com	4	1@hotmail.com	3

accounts had no activity. In these cases, it is unclear why users would create these accounts and not use them. They may be required to create accounts to participate in activities on the website that are not reflected on their profile, however, there is no available evidence as to their true purpose.

5 Discussion

5.1 Reasons for Multiple Accounts

Fake Email Addresses We found many fake email addresses in our data. We identified these by making up obviously fake addresses, generating the SHA1 hash, and searching for that hash in the list of `mbox:sha1_sum`s collected from the profiles. Aside from the blank email address discussed above, table 2 shows 20 made up addresses with many associated accounts. While these are the most popular of the fake addresses we guessed at, we found many others with only one or two associated accounts.

Different Personas Users often change information in their profiles to represent variations on their personality or life situation. It was fairly common to see users with different ages but otherwise identical profile information; we found that in 29 profiles that belonged to the same person but that had different ages. This was frequently a difference between being under 21-years-old and 21 or older. We also found variations in the profile's sexual orientation and in relationship status.

For example, Buzznet had a high population of female users in their mid-teens. It was not uncommon to see them create new profiles with new usernames when they started dating a boy, dedicating pictures and descriptions to their relationships, only to abandon that profile and create a new one when the relationship ends.

Less Common Reasons

– Sybil (pesudo)-Attacks - Some users created multiple accounts in order to have some accounts provide positive votes to content posted on the main account. This was only seen twice, but was a social network-based instantiation of the Sybil attack. While there was little malicious intent, this usage is

similar to what would be done to circumvent rating systems in more critical environments.

- Compartmentalizing - One of our hypotheses about why users had multiple accounts was due to compartmentalizing parts of their life, for example, maintaining separate accounts for their personal and professional lives. We found two `mbox:sha1_sum`s where this was the case, but it was not a common phenomenon.
- Errors - When users are not permitted to change their usernames, they register new accounts to change them. We found a few instances of users registering multiple accounts on the same day with user names that varied by one or two letters, correcting a typo.
- Groups - Occasionally groups of users shared a common `mbox:sha1_sum`. We particularly noticed this with music groups who may be using a common address for the band while each member maintains a separate profile.

5.2 Privacy Implications

One of our initial concerns was that people would be using multiple profiles on a social network to keep parts of their lives separate, specifically to keep their personal life private from their professional life. However, this was not an issue that we discovered. Most profiles on the same `mbox:sha1_sum` were very similar and when there were variations, they were minor. The large number of empty unused profiles also lessened the concern about privacy violations.

In most cases, users were extremely open in sharing intimate details of their lives in their profiles. There were open discussions of sexual activities, drug use, and personal conflicts. There were some variations between profiles in personal information (e.g. the user's sexual orientation was listed as "straight" in one profile and "bi" in another), but it does not appear that the user was trying to hide information by using multiple profiles; rather, these instances look more like teenagers experimenting with their personas.

Even in the few cases we found where users were keeping their professional and private lives separate, the "professional" context was quite casual and included personal information.

This is not to say that FOAF aggregation will not raise privacy concerns. One only need to look to the now defunct Plink as an example. The site aggregated FOAF from many sources but was forced to shut down in October 2004 after complaints from people who did not expect their data to be present on a site they did not sign up for. Thus, even when users have only a single profile, there are privacy concerns that arise from simply taking FOAF information that is freely available.

Independent of this general concern, however, there is little evidence that the merging of profiles will lead to information being aggregated that users intended to keep separate.

5.3 Preventing Incorrect Inferences

Most of the cases where Semantic Web reasoning over the FOAF profiles seems to merge accounts belonging to different people occur when users have fake addresses. This occurs when social networking websites do not perform an email verification for users to create profiles and when they allow a new user to register with an email address that's already in the system. Both of these were the case with Buzznet, resulting in the many duplicate accounts we found there.

6 Conclusions

In this paper, we looked at the implications if Semantic Web reasoning over FOAF data to merge multiple profiles of the same user. We were particularly interested in why users create multiple accounts, how they use them, and what challenges or benefits FOAF offers in dealing with this issue. To answer this question, we gathered FOAF data from all eleven social networking websites that produce FOAF files for their users, and analyzed the profiles of the five users who had the largest number of multiple profiles as well as profiles for 40 `mbox:sha1_sum`s with five associated accounts and 40 `mbox:sha1_sum`s that represented the much more common case of having only two profiles.

In all but a few cases, we found that all the accounts associated with a given `mbox:sha1_sum` represented the same user. We found many examples where people used fake email addresses, and this lead to many instances where profiles of different people were linked to the same `mbox:sha1_sum`.

Among the `mbox:sha1_sum`s where the accounts belonged to the same person, we made several observations. Frequently, the profiles were mostly unused; they had only basic profile information with no friends or posts. Users usually maintained identical information among all of their profiles. When there were discrepancies, they were usually in the user's age (perhaps to gain access to age-restricted areas of the website) or hometown. The most common reason for having multiple profiles appears to be to cultivate slightly different and / or evolving personas in an online environment. Merging the available FOAF information will lead to some inconsistent information in these cases. However, we found no instances where the merging of data would violate privacy that users tried to establish by separating their information into different accounts.

References

1. Ding, L., Zhou, L., Finin, T., Joshi, A.: How the semantic web is being used: An analysis of foaf documents. In: System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on. (2005) 113c–113c
2. Golbeck, J., Rothstein, M.: Linking social networks on the Web with FOAF: a Semantic Web case study. Proceedings of AAAI08 (2008)
3. Douceur, J.: The sybil attack. In: Peer-To-Peer Systems: First International Workshop, Iptps 2002, Cambridge, Ma, USA, March 7-8, 2002, Revised Papers, Springer (2002) 251